Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016

Zhaocheng Huang
School of Electrical Eng. and Tele.
The University of New South Wales
and Data61, CSIRO
zhaocheng.huang@unsw.edu.au

Brian Stasak
School of Electrical Eng. and Tele.
The University of New South Wales
and Data61, CSIRO
b.stasak@student.unsw.edu.au

Ting Dang
School of Electrical Eng. and Tele.
The University of New South Wales
and Data61, CSIRO
ting.dang@unsw.edu.au

Kalani Wataraka Gamage
School of Electrical Eng. and Tele.
The University of New South Wales
and Data61, CSIRO
kalani.watarakagamage@unsw.edu.au

Phu Le
School of Electrical Eng. and Tele.
The University of New South Wales
Sydney NSW 2052 Australia
phule@unsw.edu.au

Vidhyasaharan Sethu
School of Electrical Eng. and Tele.
The University of New South Wales
Sydney NSW 2052 Australia
v.sethu@unsw.edu.au

Julien Epps
School of Electrical Eng. and Tele.
The University of New South Wales
and Data61, CSIRO
j.epps@unsw.edu.au

ABSTRACT
Within the field of affective computing, human emotion and disorder/disease recognition have progressively attracted more interest in multimodal analysis. This submission to the Depression Classification and Continuous Emotion Prediction challenges for AVEC2016 investigates both, with a focus on audio subsystems. For depression classification, we investigate token word selection, vocal tract coordination parameters computed from spectral centroid features, and gender-dependent classification systems. Token word selection performed very well on the development set. For emotion prediction, we investigate emotionally salient data selection based on emotion change, an output-associative regression approach based on the probabilistic outputs of relevance vector machine classifiers operating on low-high class pairs (OA RVM-SR), and gender-dependent systems. Experimental results from both the development and test sets show that the RVM-SR method under the OA framework can improve on OA RVM, which performed very well in the AV+EC2015 challenge.

Keywords
Depression classification; dimensional emotion prediction; token word selection; relevance vector machine; output-associative fusion; annotation delay compensation; gender dependence; multimodal fusion.

1. INTRODUCTION
Affective computing relates to the study and development of automated emotion comprehension. It can be applied to interpret an individual's emotional state, as well as identifying abnormal behavior originating from disorders or disease. For example, healthcare professionals and engineers are considering new reliable methods to identify and monitor depression. While the effect of depression on verbal communication has been researched over many decades [1], [2], it is only recently that advances in automatic affective computing (e.g. speech, facial analysis, body language) have been successfully applied to help identify and predict levels of depression. Affective computing analysis of depression has its advantages. For instance, considerable time is spent on training new qualified mental health experts and providing adequate evaluations for the sizeable and growing number of patients worldwide with depression. Visual and audio processing of individuals can quickly provide discrete temporal information and discover social-behavioral patterns at a cohort level that may be missed by a mental health clinician, especially if only one modality is available.

Another important field within affective computing is the development of inexpensive machines capable of automatically recognizing people's emotional status using non-invasive methods. Continuous analysis of affect, such as arousal and valence, has been shown to be effective in capturing subtle transitions in emotions in naturalistic environments. This has motivated an increasing number of developed systems for continuous prediction of affect dimensions.

In continuous affect and depression research, multi-modal behavioral analysis using audio, visual, and other non-invasive physiological signals have demonstrated potential for use as a tool for clinicians. Moreover, motivation stemming from the annual Audio-Visual Emotion Challenges and Workshops (AVEC) [3], [4], has led to more insightful automated affective computing designs, resulting in better overall performances with these two tasks and understanding of human behavior. Similar to the challenges presented in the last few years, the AVEC 2016 [5] provides a platform-benchmark for developing emotion prediction and depression recognition systems.

The research herein describes and tests systems as entries for the affect and depression AVEC 2016 sub-challenges [5]. For the depression sub-challenge, our investigations focus on spectral centroid frequency-based vocal tract coordination features

AVEC'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2988257.2988265

(Section 3.2) and 'thin slice' acoustic analysis of token words (Section 3.1); for the emotion sub-challenge, we investigate a two–stage prediction approach based on probabilistic outputs from Relevance Vector Machine classification of ranked class-pairs (Section 4.2), and data selection (Section 4.3) from a novel emotion change perspective.

2. RELATED WORK

2.1 Depression

Observed speech-language behaviors in depressed individuals often include a combination of the following: a reduction in pitch, prosody, loudness, and rate of speech [2]. Furthermore, in severely depressed individuals motor incoordination (e.g. motor retardation) and/or word retrieval difficulty is often also a behavior exhibited [2], [6]. Spectral, prosodic, voice quality and glottal acoustic features have been investigated for the detection and prediction of depression levels in individuals' speech. The most commonly used acoustic features include pitch, formants, formant bandwidths, intensity, harmonic-to-noise ratio, shimmer, jitter, and overall rate of speech [2], [7]. Recent investigations into the effects of depression on speech production have indicated that combinations of these types of acoustic features can provide strong depression recognition performances [8]. Moreover, acoustic features fused along with other modalities (e.g. visual) have shown robustness in depression recognition performances across speakers [9].

In other speech processing applications, such as speaker and language identification, text-dependent analysis has shown performance improvements over using larger portions of data which constitute greater phonetic variability [10]. Previous studies have indicated that 'thin slice' data selection of a single word, phrases, or even disfluencies can generate competitive emotion/depression classification performances [11], [12]. For example, in [11], using only a small beginning portion of spoken phrases rather than all phrases, superior depression classification results with significantly less data was demonstrated.

It has been shown that acoustic features perform differently between males and females [13]–[15], but the investigation of gender-dependent modeling for depression recognition has only been described in a few studies. For instance, in [13], researchers discovered stronger correlation between depression level scores and prosodic features such as pitch and formants for male speakers than for female speakers. Other research has looked specifically at depression classification per gender [16]–[18] and found that gender-dependent models are advantageous in depression classification performances. These studies focused specifically non-verbal gender-based depression behavioral differences in the visual modality. While [18] found that both genders had repressed non-verbal expressiveness when compared to healthy speakers, the depressed females presented more socially interactive behaviors than the depressed males. In [16], the use of gender-dependent models for depression classification was also advocated due to the high-accuracy of automatic gender identification based on audio and/or visual information.

In previous AVEC depression challenges, Vocal Tract Coordination (VTC) features extracted from formants and del-mel-cepstra features produced by far the state-of-the-art performances in Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for predicting Beck Depression Inventory (BDI) scores in depression [19], [20]. Features of this kind were originally proposed in [21], where they were used to capture changes in electro-encephalogram (EEG) spatiotemporal correlation structure for seizure prediction. In [19], it was showed that the VTC features can represent the changes in vocal tract coordination associated with depression and consequently can improve depression prediction from a number of other modalities.

To date, the relative performance advantages of these three suggested methods – token word-based data selection, gender-dependent modelling and VTC features – remain unknown.

2.2 Emotion

The AVEC 2016 challenge [5] adopted the same data partitions of the RECOLA corpus [22] as AV+EC 2015. Despite this, with a simpler version of the back-end, i.e. linear Support Vector Regression (SVR), the baseline has been raised significantly to be far more competitive compared with that of AV+EC 2015. This is largely because of the optimization of window sizes for feature extraction and delay compensation in the gold standard for each single modality, and for arousal and valence respectively. In addition, pre-processing such as speaker-wise and global *z* normalization, and post-processing such as centering and scaling have been considered as parts of the baseline systems. Linear regression at decision level further combines multiple modalities to achieve test scores in Concordance Correlation Coefficients of 0.682 for arousal and 0.638 for valence [5]. However, it is believed that there still exists a range of alternative approaches for a more robust and effective multimodal prediction system. Approaches investigated herein include data selection for emotionally salient segments and exploitation of uncertainties of RVM outputs.

One potential limitation of current continuous emotion prediction systems is to treat all input features as being equally emotionally-related. This may not be a good assumption in general, because within-utterance variation and emotional saliency in features are disregarded. In other words, there may exist emotionally salient segments within each utterance that can contribute more to emotion prediction. This aspect has been explored less compared with emotion classification systems, where e.g. segment selection based on specific phonemes or phoneme classes has been found more conducive to emotion classification [23]–[25]. Another investigation [26] into selecting informative segments within utterances examined various data selection strategies based on classifier agreement.

The Relevance Vector Machine (RVM) [27] employs a probabilistic Bayesian framework to achieve a sparse representation for regression and classification, and is a relatively new regression model for predicting affect dimensions [28]. RVM makes the prediction based on the form [27]:

$$y(\boldsymbol{x}_*, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_*) + \boldsymbol{\epsilon} \quad (1)$$

$$P(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{K} N(\boldsymbol{w_i}|0, \alpha_i^{-1}) \quad (2)$$

where $\boldsymbol{x}_*$ represents a *k*-dimensional feature vector, and $\boldsymbol{w} = [w_1, \cdots w_K]^T$ are the weights of each feature dimension. The sparsity is enforced in the weights by introducing a zero-mean Gaussian prior controlled by $\boldsymbol{\alpha} = [\alpha_1, \cdots \alpha_K]^T$, as seen in (2). $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T$ denotes the training noise terms.

A more effective framework for building emotion prediction systems from RVM is called Output-Associative (OA) RVM [28], formulated as:

$$y_t^a = (\boldsymbol{\omega_a})^T \boldsymbol{\phi}(\boldsymbol{x}_*) + (\boldsymbol{\varphi_a})^T (\tilde{\boldsymbol{y}}_t^a) + (\boldsymbol{\psi_a})^T (\tilde{\boldsymbol{y}}_t^v) + \boldsymbol{\epsilon} \quad (3)$$

$$y_t^v = (\boldsymbol{\omega_v})^T \boldsymbol{\phi}(\boldsymbol{x}_*) + (\boldsymbol{\varphi_v})^T (\tilde{\boldsymbol{y}}_t^a) + (\boldsymbol{\psi_v})^T (\tilde{\boldsymbol{y}}_t^v) + \boldsymbol{\epsilon} \quad (4)$$

where $\tilde{\boldsymbol{y}}_t^a$ and $\tilde{\boldsymbol{y}}_t^v$ are sets of temporal arousal and valence predictions at frame *t*. With the inclusion of dependencies between 1) temporal predictions; 2) affect dimensions; 3) predicted outputs and input features, this framework is beneficial for fusion and previously showed robustness for a range of system

settings [29]. An advantage of RVM is that it offers probabilistic outputs representing uncertainty for each prediction. A similar probabilistic output can be obtained from RVM-based classifiers, which can further be exploited to improve system performances, although this has yet to be investigated in the literature.

Female and male speakers may express their emotions differently which may cause side-effects on speech based emotion recognition systems [30]. Attempts at reducing the gender effects involve training gender-dependent systems or penalizing the gender variability during modeling - both [31] and [32] reported superior performances in emotion classification using these two approaches, respectively. In contrast to emotion classification, emotion prediction systems exploiting gender information considered have been explored relatively less. In addition to the difference in emotion expression, we speculate that there may be some differences in the emotion perception of annotators, who may tend to rate female and male speakers differently. Motivated by this, we studied gender-dependent emotion prediction systems.

3. DEPRESSION INVESTIGATION

3.1 System Overview

The following three sub-systems are proposed for classifying depression: audio, video and token word systems, as seen in Figure 1. In the audio system, we extracted short-term acoustic features. These features were then used to calculate the VTC features under different sliding window sizes before classification modeling. For the video system, we used two sets of *FACET* features (emotion traces and Action Units) and four sets of *OpenFace* features exclusive of the HoG features due to their high computational complexity and relatively poor performances. These features were directly or indirectly (VTC feature extraction) used for classification modeling. Training and Testing were performed on a per-frame basis and majority voting was applied to generate one output per file. For the token word system, speech segments for specific words were identified based on transcripts that include time stamps for spoken words. Acoustic features from these words were then extracted for training a model for classification.

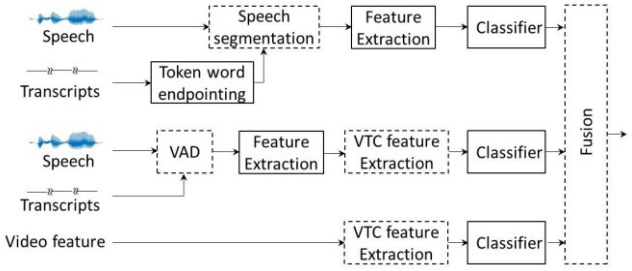

Figure 1: *Overview of depression systems*

Train and test segments in the token word system are expected to have similar phonetic content, and to be of sufficient duration so as to capture reasonably long term acoustic context. This can help reduce phonetic variability.

3.2 Vocal Tract Coordination Features

Motivated by [19] and [20], we examined the VTC features extracted from four sets of short-term acoustic features and six sets of video features for the depression classification task. The features were 16-dimensional MFCC, 16-dimensional delta MFCCs, 13-dimensional Spectral Centroid Frequencies (SCF) and 13-dimimensional Spectral Centroid Magnitudes (SCM) [33]. These four feature sets capture different characteristics of the spectral envelope. For example, the SCF feature estimates the frequency 'centre of gravity' of the speech spectrum within individual sub-bands, hence characterizing the distribution of the sub-band spectral energy [33]. This feature is a formant-like feature as it gravitates towards the location of formant frequency in each sub-band. On the other hand, the SCM feature estimates the weighted average magnitude spectrum in the sub-band. VTC features are sensitive to changes in the temporal delays between 'channels', i.e. different parameter contours, within a particular feature set. They are thought to be sensitive to psychomotor disturbances because during disturbances muscular control behavior may become incoordinated, resulting in atypical or unsynchronized parameter evolution over time.

During VTC feature extraction, auto and cross-correlation were calculated within and between channels. A certain number of correlation points were selected from the auto and cross-correlation sequences and their lagged versions to form a channel-delay correlation matrix, whose eigenspectra were then calculated. This process was repeated for multiple time-scales, i.e. different spacing between correlation points, and all eigenspectra were concatenated to form a new feature vector. PCA was further applied to eliminate the highly correlated features. A more detailed description can be found in [20]. The channel-delay correlation matrix is capable of capturing not only variations in correlation at different time scales, i.e. high-frequency changes with small spacing and low-frequency changes with large spacing, but also coherence or interaction between different channels, which may help distinguish between healthy and depressed speakers.

4. EMOTION INVESTIGATION

4.1 System Overview

The AVEC 2016 emotion sub-challenge provides feature sets from audio, video and physiological (ECG and EDA based) signals. The feature sets include an 88-dimensional acoustic feature set: the *Extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [34], two sets of facial based video feature sets (168-dimensional appearance features and 632-dimensional geometric features) and physiological features. For each feature set, window sizes for feature extraction were optimized for arousal and valence respectively. Further details on provided features can be found in [5].

The proposed systems utilize the eGeMAPS features, the two sets of video features and a set of 650-dimensional audio features which contains five statistical functionals, i.e. mean, standard deviation, min, max and max-min range calculated over 130-dimensional ComParE 2013 low level descriptors using the *openSMILE* toolkit [35]. These functionals are calculated within a 2-second window in every 40 milliseconds, referred as to ComParE 2013. The physiological features were omitted, as they yielded marginal improvements during the system development.

An overview of the final emotion prediction system is illustrated in Figure 2. The four selected feature sets were fused to generate 1538-dimensional features and delay compensation was carried out on both training and test data. The training data were used for regression modeling with OA RVM or RVM-Staircase Regression. The proposed data selection method based on emotion changes was applied to training data; refer to Section 6.2.3 for further detail.

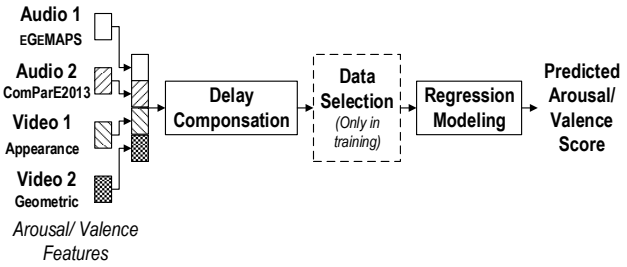

Figure 2: *Overview of emotion prediction system (with optional data selection from training data)*

4.2 RVM Staircase Regression (RVM-SR)

This approach is motivated by Gaussian Staircase Regression (GSR), which was first proposed for depression prediction [19]. In the GSR approach, data corresponding to intervals of the rating scale were grouped into several pairs of low-high classes, and the log mean likelihood ratio (LMLR) between the low and high partition was calculated. The LMLR from each low-high class pair was then used in regression modeling, to predict depression BDI scores. Based on the same framework, RVM-SR used an ensemble of RVM *classifiers* to model the class boundaries associated with low-high class pairs. The probabilistic outputs from each of the RVM classifiers were used as features for training a regression model at the second stage. The assumption made in the previous application of RVM-SR to depression score prediction [36], was that probabilistic outputs of the RVM classifier reflect beliefs in how strongly an utterance corresponds to a certain region of depression BDI scores. This may hold true for emotion ratings, which motivated investigation of RVM-SR for emotion prediction tasks herein.

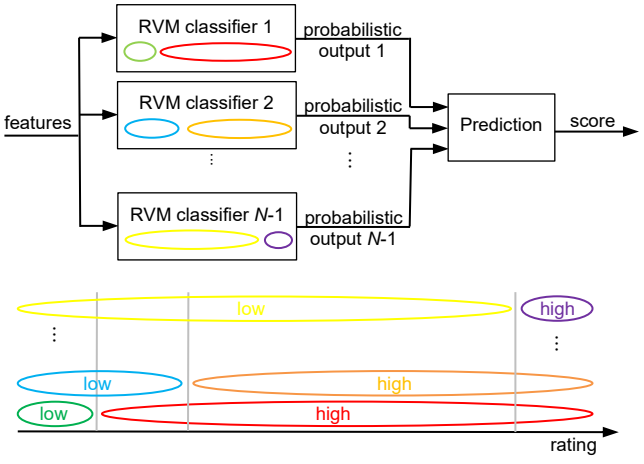

Figure 3: *Overview of RVM Staircase Regression (RVM-SR) approach, showing how pairs of low-high classifiers are built upon intervals of the rating dimension, after* [36].

Similarly to RVM regression in (1) and (2), the RVM classifier learns a set of sparse weights for the most relevant features during training:

$$y(\boldsymbol{x}_*, \boldsymbol{w}) = \boldsymbol{w}^T \psi(\boldsymbol{x}_*) \quad (5)$$

where $y(\boldsymbol{x}_*, \boldsymbol{w})$ follows a Bernoulli distribution and is combined with a sigmoid function $\varsigma\{y\} = \frac{1}{1+e^{-y}}$ to represent the posterior probability of class membership [27].

$$p(c_*|\boldsymbol{x}_*, \boldsymbol{w}) = \varsigma\{y(\boldsymbol{x}_*, \boldsymbol{w})\} \quad (6)$$

Both the arousal and valence dimensions were divided into N classes (based on the distribution of their ratings, where percentiles are evenly divided), with a view to compare adjacent pairs of low-high classes, as seen in Figure 3. One method proposed for scoring the similarity of a feature under test to the low or high class is to use the probabilistic output of an RVM classifier $\{p(c_n|\boldsymbol{x}_*, \boldsymbol{w^n})|\ n \in [1,2,\dots,N-1]\}$, resulting in N-1 probabilities for the adjacent low-high class-pairs from which the rating is predicted. This has shown fairly good results for depression prediction [36], however it has not been applied to emotion prediction so far.

4.3 Data Selection based on Emotion Change

The data selection that was evaluated in this work is based on the assumption that segments corresponding to change in emotion are more salient than segments within which there is little or no change. To illustrate the various segments considered for selection, in Figure 4 partition C contains all the frames where emotion ratings change in addition to all frames whose ratings remain unchanged for less than L frames. Partition B contains all the frames before emotion ratings change at the beginning of every file, and the second half of all frames whose ratings remain unchanged for more than L frames. Similarly, partition A contains all the frames after the last change frame of every file and the first half of all frames whose ratings remain unchanged for more than L frames, where L is the minimum number of frames considered as non-change frames (i.e. B and A). Non-change frames were divided into B (before) and A (after) for exploring differences before and after emotion change frames. L was introduced to provide more continuity for C, as seen in Figure 4 With L = 1, partitions B, C and A account for 16.86%, 63.18%, and 19.96% respectively for arousal while those corresponding measures for valence are 19.36%, 58.10%, and 22.54%.

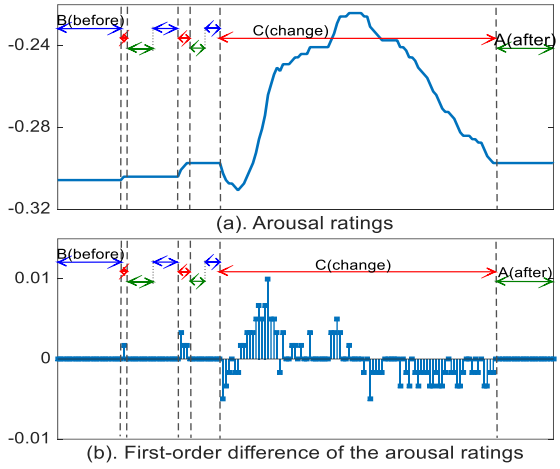

Figure 4: *Example arousal rating vs. time, showing the definitions of the different partitions relative to emotion changes*

Since the first-order differences may fluctuate due to annotator uncertainty or tremble, partitioning training data based on those first-order differences may be unreliable for emotion changes. To resolve this, we smoothed the first-order differences by applying a low-pass filter, whose length was empirically set to 15 and 30 frames. Then a threshold was applied to detect a frame with the largest emotion change compared to its previous frame.

5. EXPERIMENTAL CONDITIONS

5.1 Database

The depression sub-challenge adopts the Distress Analysis Interview Corpus (DAIC) [6][1]. It was originally designed to investigate clinical communication behaviors and assisted human-

[1] Quality control listening per utterance was completed to check time-stamp accuracy (some transcriptions had time-stamp errors).

computer commutative dialog. This database provides a fixed set of prompts; a large group of speakers, 189 speakers in total, which are divided into training (107 speakers), development (35 speakers) and test (47 speakers) partition; high-quality close-talking microphone recordings; natural speech; and PHQ-8 evaluations along with scores per speaker. The PHQ-8 has an interval scale from 0 to 24 and larger scores indicate greater depression severity. The DAIC also includes phrase-level transcripts with beginning/ending time markers, making locating single token words possible with minimal error. Further details of the DAIC transcription conventions could be found at [6].

The AVEC 2016 emotion sub challenge was assessed on part of the *Remote Collaboration and Affective Interaction* (RECOLA) corpus [22], containing speech of 27 subjects which were evenly divided into training, development (devel.) and test partitions. The database comprises synchronously recorded multimodal signals, i.e. audio, video, and physiological signals for the first 5 minutes of each conversation. Arousal and valance emotional attributes during this period were rated by 6 gender balanced raters in 40ms time steps, resulting in 7501 pairs of affective score per file.

5.2 Key Experimental Settings

In depression classification systems, the token words (primarily 'filler' words) used for experiments were: *hmm*, *mhm*, *no*, *uh*, *so*, *mm*, *umm*. These words were chosen as they commonly occur throughout conversations, and performed well relative to other token words with relatively constrained phonetic content. The token words *hmm*, *mhm*, *no*, and *uh* combined were spoken by 95% of speakers from the training, 100% of speakers from the development and 100% of speakers from the test sets. Roughly 5% of training data were omitted due to transcript time marker errors. Since token words present a shortage of training data for a classification model, we trained a linear SVR model with complexity coefficients selected from $[10^{-5}, 10^{0}]$ to predict PHQ-8 scores, which were then thresholded into depressed or non-depressed. The threshold was empirically set to $T = 7$ due to consistent classification performances across various sets of features. This study adopted the same settings for the VTC feature extraction from audio and video features as per [20], except excluding the log power and entropy of the channel-delay covariance matrix due to no observed improvement. That is, 4 timescales with spacing of 1, 3, 7, 15 frames respectively. The number of correlation points was set to 15. For audio features, we performed Voice Activity Detection based on the transcription. VTC was applied to four sets of short-term acoustic features (Section 3.2) and provided video features.

In emotion prediction systems, speaker-wise preprocessing such as [0, 1] scaling and *z*-normalization were implemented for each modality. We retained whichever of the methods gave better performance, and observed that speaker-wise *z*-normalization generally performed poorer in terms of CCC. We applied the speaker-wise [0, 1] scaling to appearance-based video features (arousal and valence) and geometric-based video features (arousal only). For post-processing, we applied centering and scaling as in the baseline paper [5] for arousal prediction within all systems. During system development, we trained on all training data and tested on all development data. Training data were scaled into the range [0, 1] and scaling coefficients were used to normalize test data. For RVM, the back-end for emotion prediction systems, the only parameter to be tuned was the iteration number, which was selected among {10, 30, 50, 70}. In RVM-SR, the distribution of ratings on training data was evenly partitioned into 10 classes for arousal and 20 classes for valence, selected empirically, which means that the number of pairs of adjacent low-high partitions for arousal and valence were 9 and 19 respectively. The window sizes for OA RVM were set to 201 frames for arousal and valence.

6. SYSTEM DEVELOPMENT

6.1 Depression Systems

6.1.1 VTC vs Token Words

A comparison between token words versus VTC features was completed using MFCC, delta MFCC, SCM, and SCF features, as shown in Table 1. In both systems, training and testing were performed on a frame-basis, and outputs from test frames were based on either majority voting (VTC) or average (token words). The best VTC entire utterance baseline feature was the SCF feature, which attained an F1 score of 0.50(0.74). Results for the 4-best token words ('*hmm*','*mhm*','*no*','*uh*') raw features are shown in Table 1. Note that SCF features from the 4-best token words outperformed the VTC and the AVEC 2016 audio/visual development baseline in F1 score. Additionally, the 4-best token word SCM and SCF functionals were explored and performed similarly to the VTC and AVEC 2016 audio baselines. 7-best token words were evaluated as well; however, this did not outperform the 4-best. As previously hinted at in [14], it was believed that thin slice token word selection could potentially do better across all the development than using 100% of each utterance, and this hypothesis seems to be supported in Table 1, where token words are generally better in average F1 scores.

Table 1: *F1 scores for VTC audio features vs token words*

Front-end features	VTC (Audio)	Token words
16-dim MFCC	0.43(0.72)	0.33(0.89)
16-dim ΔMFCC	0.36(0.71)	0.22(0.89)
13-dim SCF	0.50(0.74)	**0.59(0.87)**
13-dim SCM	0.50(0.74)	0.36(0.88)
Combined	0.42(0.70)	0.40(0.84)

6.1.2 Whole File vs Thin Slice

This section compares 88-dimensional eGeMAPS features [34] extracted at different time scales: the whole utterance, 3 second sliding window with 1 second overlapping, and token word segments. Interestingly, the best performances were achieved using the least amount of data; a significant and promising 0.54(0.92) on the development set.

Table 2: *Comparison of eGeMAPS features extracted at different time scales.*

Systems	F1 score	Precision	Recall
Whole file	0.27(0.80)	0.25(0.81)	0.29(0.79)
3s (1s) sliding window	0.44(0.92)	1.00(0.85)	0.29(1.00)
Token words-4	**0.54(0.92)**	0.75(0.87)	0.43(0.96)
Token words-7	0.36(0.88)	0.50(0.84)	0.29(0.93)

6.1.3 VTC for Video Features

This section compares video features and their VTC features. The sliding window sizes for VTC extraction were optimized on the development set among {10, 20, 40, 80, 120} seconds, except that only 120s was used for the CLM-2D and CLM-3D features because of their high dimensionality, causing rather long computation time. The VTC features with optimized window size outperformed raw features in general. The CLM-2D and CLM-3D achieved higher F1 scores compared with their VTC features.

Table 3: *Comparison of VTC features extracted from video features and raw video features in F1 score*

Front-end features	VTC (Video)	Video
FACET emotions	0.38(0.81)	0.30(0.56)
FACET AUs	0.31(0.84)	0.17(0.60)
CLM-2D	0.29(0.82)	0.50(0.70)
CLM-3D	0.42(0.78)	0.44(0.81)
CLM-Gaze	0.40(0.76)	0.33(0.65)
GLM-Pose	0.38(0.81)	0.44(0.81)

6.1.4 Decision Level Fusion

This section examines performances for different combinations of previous sub systems. Decision level fusion was achieved using logistic regression. The first set of results combined four individual VTC systems for four sets of acoustic features. This yielded higher performances than a single VTC-audio system. The best audio and video subsystems were combined, yielding further improvements. Different combinations of audio and video systems were trialed, but none of them outperformed the VTC-SCF + CLM-2D system, shown in Table 4. In addition, inclusion of token word system at decision level did not favor classification.

Table 4: *Performances for decision-level fusion of the best audio, video and token word systems.*

Systems	F1 score	Precision	Recall
VTC-Audio fused	0.53(0.82)	0.42(0.91)	0.71(0.75)
VTC-SCF + CLM-2D	0.59(0.87)	0.50(0.92)	0.71(0.82)
VTC-SCF + CLM-2D + Token words-4	0.59(0.87)	0.50(0.92)	0.71(0.82)

6.1.5 Gender-dependent Systems

The best system from the previous section, i.e. the token word subsystem using the SCF features, was used to further examine the effect of gender difference in depression classification, shown in Table 5. The starting point was to train and test on only either female or male data, producing two sets of improved performances as shown in Table 5. This was achieved with optimized models and threshold T for each. The results were further combined to generate final gender-dependent system results. As shown in Table 5, the results indicate that separate gender-based systems for depression classification analysis perform better than gender-agnostic systems, especially in the case of male speakers. Similar and consistent performances were observed using EGEMAPS features with the same system configurations.

Table 5: *4-best token word gender-dependent depression classification performances using raw SCF features*

Systems	F1 score	Precision	Recall
Gender-independent (T=7)	0.59(0.87)	0.50(0.92)	0.71(0.82)
Female (T=6.75)	0.60(0.86)	0.43(1.00)	1.00(0.75)
Male (T=7)	0.73(0.86)	0.57(1.00)	1.00(0.75)
Gender-dependent	0.67(0.86)	0.50(1.00)	1.00(0.75)

6.2 Emotion Systems

As delays are crucial to the accuracy of emotion prediction systems, the first development priority was to optimize delays for each modality. The delay value was optimized in CCC on development set within the range [0, 6] seconds with a 0.4 second increment. The optimized delay values (as seen in Table 6) were adopted throughout all experiments. Compared with single modality performances in the baseline paper [5], the provided EGEMAPS feature set performed slightly poorer. However, the video-appearance features provided superior performances for arousal (0.615 vs 0.483) and valence (0.530 vs 0.470). Similar improvement can be seen for arousal using video-geometric features for arousal (0.467 vs 0.379). These improvements were from speaker-wise scaling on features before training and testing.

Table 6: *Performances and optimized delay values for single modality using RVM*

	Arousal		Valence	
	CCC	Delay	CCC	Delay
EGEMAPS	0.750	2.8	0.396	2.4
ComParE 2013	0.750	3.2	0.361	2.8
Video-appearance	0.615	2.4	0.530	2.8
Video-geometric	0.467	4.0	0.571	3.2

6.2.1 RVM vs RVM-SR

RVM has shown solid emotion prediction performances [29], but systems considering its probabilistic outputs have been investigated relatively less. As seen in related work on depression prediction [36], regression approaches based on probabilistic outputs of RVM classifiers operating on low-high class pairs perform well. We applied this idea to continuous emotion prediction. In RVM-SR, to partition the data into different low-high pairs, we evenly divided percentiles from the training ratings (i.e. arousal and valence) based on which thresholds are selected. Data was grouped into 10 partitions based on arousal scores and 20 partitions based on valence scores to give 9 and 19 corresponding low-high class pairs and used these to train RVM classifiers (Figure 3). The sets of low-high class posteriors from these classifiers were used to train a RVM regression model.

Table 7 compares RVM and RVM-SR. In two systems, features from 4 modalities were concatenated at the feature level. The RVM system was outperformed by RVM-SR, which exploits the probabilistic output property of conventional RVM. These low-dimensional probabilities are found to be informative for emotion prediction tasks, which can be further incorporated within the OA-RVM framework. It was also observed that RVM-SR performs relatively poorly on a per-modality basis, but better than RVM when features from 4 modalities are combined.

Table 7: *Comparison of RVM and the two-stage RVM-SR. Note that the dimensionalities of features were reduced to 9- and 19-dim within RVM-SR, compared with 1538-dim features within RVM.*

Systems	Arousal	Valence
RVM	0.675	0.586
RVM-SR	0.770	0.640

6.2.2 OA RVM vs OA RVM-SR

The output-associative RVM (OA RVM) framework is effective for continuous emotion prediction [28]. This section compares RVM and RVM-SR under the OA framework. In OA RVM-SR, probabilistic outputs from RVM-SR are included within the OA matrices, which include temporal arousal and valence predictions and input features for further training a regression model. This enables more information about the low-high classes to be considered during regression modeling. However, results shown in Table 8 indicate that the proposed OA RVM-SR performed slightly poorer for arousal and valence, which is presumably because the probabilistic outputs are low-dimensional compared with the original OA matrices, which are more than 2000 dimensions. Since RVM generates models with sparse weights for features, the low-dimensional information may not favor regression in this context.

Table 8: *Comparison of RVM and RVM-SR under the OA fusion framework*

Systems	Arousal	Valence
OA RVM	0.857	0.695
OA RVM-SR	0.855	0.642

6.2.3 Data Selection based on Emotion Changes

To identify emotional saliency in speech and video, this section examines the proposed emotion change based data selection method, as described in Section 4.3. Table 9 shows that the change (C) and non-change (B and A) partitions produced slightly lower CCCs than using all data. This may be in part due to noises and arbitrary fluctuations in the first-order differences of arousal and valence, from which data were partitioned into B, C and A. Therefore, we applied a mean filter of length *W* in frames to smooth the first-order differences. In addition to *W*, there were two parameters used to characterize partition C: threshold *T* to select large changes, and regions *R* in frames around large emotion changes included as C. [*W*, *T*, *R*] were set to [30, 0.0012, 5] and [15, 0.0016, 5] to partition C for training. With smoothing, training on only C (change frames) achieved better performances for valence.

Table 9: *Comparison of performances using all and subset of training data*

Partitions	Arousal	Valence
B+C+A	0.859	0.696
C	0.809	0.647
B	0.814	0.681
A	0.803	0.676
C [15, 0.0016, 5]	0.804	0.664
C [30, 0.0012, 5]	0.812	**0.707**

6.2.4 Gender-dependent Systems

This section examines effects of gender variability on emotion prediction systems under the OA RVM framework. Three gender-dependent systems were developed: *Female*, *Male* and *Combined*. For *Female* and *Male* systems, models were trained and tested using Female only or Male only data, whereas in the *Combined* system, predictions from *Female* and *Male* systems were combined to generate results as shown in Table 10. The gender dependent model yielded poorer performances for emotion prediction, likely because there is less data for training.

Table 10: *Comparison of gender-dependent and gender-independent systems*

Systems	Arousal	Valence
Gender-independent	0.855	0.654
Female	0.851	0.622
Male	0.808	0.597
Combined	0.831	0.612

7. AVEC 2016 CHALLENGE RESULTS

7.1 Depression

For the challenge test submission, depression classification results using the 4-best token words SCF features, gender-dependent models, and decision-level fusion of VTC-SCF and VTC-MFCC had performances below the baseline [6]. In all the systems, training and development data were merged for training. The test results in Table 11 clearly show lower depressed identification for all three methods in clear contrast to development results shown previously in Sections 6.1.1 and 6.1.2. For the 4-best token words, the lower performance is possibly attributable to selecting a balanced threshold and complexity coefficient. As supporting evidence, an improvement was recorded by adjusting the latter to a higher value (as indicated in Table 11 by an '*'). However it is also likely that, despite care with feature dimensionality, overfitting was an issue in these experiments. In general, it was observed that there was little similarity between system configurations which performed well on the development and system configurations which performed well on the test set. This suggests that the database contains considerable variability.

Table 11: *Comparison of AVEC 2016 test set for depression sub-challenge*

Systems	F1 score	Precision	Recall
Baseline [5]	0.50(0.90)	0.60(0.87)	0.43(0.93)
Token word-4 SCF	0.14(0.85)	0.20(0.81)	0.11(0.90)
Token word-4 SCF*	0.30(0.81)	0.27(0.83)	0.33(0.79)
Token word-4 SCF GD	0.20(0.78)	0.18(0.81)	0.22(0.76)
VTC-SCF+VTC-MFCC	0.17(0.81)	0.12(0.72)	0.33(0.40)

7.2 Emotion

Three emotion prediction systems were submitted as affect sub-challenge entries. In all systems, training and development data were merged for training. All three systems employed an iteration number of 10 for arousal and 30 for valence, optimized on development set.

The first system was an *OA RVM* system. The OA window size was set to 201 frames and 21 frames optimized on the development set for arousal and valence. The second system, *OA RVM-SR,* included probabilistic outputs of RVM classifiers. The partition number was set to 10 and 20 for arousal and valence. In the system, RVM-SR was trained on each modality, and probabilistic outputs from each modality were included into OA matrix for regression modeling. The third system employed the proposed data selection under the OA RVM framework. Parameters [*W, T, R*] to partition C were set to [15, 0.0016, 5], chosen because of their consistent performances in 2-fold cross validation. This resulted in the use of 57.61% and 38.18% of training data for arousal and valence. For all submissions, we achieved improved performances over the baseline for arousal, but poorer for valence.

Table 12: *Comparison of AVEC 2016 test set for affect sub-challenge*

	Arousal	Valence
Baseline [5]	0.682	0.638
OA RVM	0.770	0.533
OA RVM-SR	0.770	0.545
OA RVM with data selection	0.728	0.515

8. CONCLUSIONS

In this paper, experiments on and submissions to the AVEC 2016 challenge were reported, with the primary focus on the audio subsystem. For the continuous emotion system, the OA framework showed consistently superior performances, confirming the effectiveness of introducing temporal dependencies of emotion attributes. The introduction of a prediction approach based on probabilistic outputs of Relevance Vector Machine (RVM) classifiers operating on low-high class pairs (RVM-SR), which is novel in the emotion prediction context, provided a significant improvement over RVM prediction results on the development set and also outperformed RVM on the test set in a novel configuration under the output-associative (OA) framework. These results are broadly in line with other methods based on predictions from pairwise comparisons [20], [36]. Overall stronger test set results for arousal than valence were expected, with the focus on the audio subsystem, and the OA

RVM-SR prediction result for arousal provided 10% higher accuracy than that of the baseline.

Novel explorations into thin slice token word depression classification and gender-specific depression modelling expanded upon previous research findings. Results from investigating these methods show that data selection has a significant impact on depression classification performance, which was strongly positive on the development set, with significant gains over the development set baseline results. Moreover, for similar acoustic data (or even in other modalities), minimal data were required to achieve high classification accuracy. The token word approach performed more poorly for the test set, and there were insufficient test submissions to understand the reasons behind this (possibly due to the issue of balancing the classifier output).

Experiments on gender dependency were conducted on both emotion and depression systems, which provided higher performances for depression classification on the development set, but no improvement was seen for emotion prediction systems.

Future work will explore a larger set of token words, helping determine whether a combination of reduced phonetic variability and word type (or phrase location) contributes to depression classification. Also, the gender depression modeling will be further investigated; analyzing a greater number of gender-specific features across modalities.

9. ACKNOWLEDGEMENT

Data61, CSIRO is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

10. REFERENCES

[1] Newman, S. and V. Mather, "Analysis of spoken language of patients with affective disorders," *American Journal of Psychiatry*, vol. 94, no. 4, pp. 913–942, 1938.

[2] Cummins, N., S. Scherer, et al., "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015.

[3] Valstar, M., B. Schuller, et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on AVEC, ACM MM*, 2014, pp. 3–10.

[4] Ringeval, F., B. Schuller, et al., "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proceedings of the 5th International Workshop on AVEC, ACM MM*, 2015, pp. 3–8.

[5] Valstar, M., J. Gratch, et al., "AVEC 2016 – Depression , Mood , and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on AVEC, ACM MM*, 2016.

[6] Gratch, J., R. Artstein, et al., "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2014.

[7] Asgari, M., I. Shafran, et al., "Inferring clinical depression from speech and spoken utterances," in *2014 IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[8] Cummins, N., J. Epps, et al., "An Investigation of Depressed Speech Detection: Features and Normalization.," in *INTERSPEECH*, 2011.

[9] Cummins, N., J. Joshi, et al., "Diagnosis of depression by behavioural signals," in *Proceedings of the 3rd ACM international workshop on AVEC, ACM MM*, 2013, pp. 11–20.

[10] Hébert, M., "Text-dependent speaker recognition," *Springer handbook of speech processing*, pp. 743–762, 2008.

[11] Alghowinem, S., R. Goecke, et al., "Detecting depression: a comparison between spontaneous and read speech," in *ICASSP*, 2013, pp. 7547–7551.

[12] Moore, J., L. Tian, et al., "Word-level emotion recognition using high-level features," in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2014, pp. 17–31.

[13] Hönig, F., A. Batliner, et al., "Automatic modelling of depressed speech: relevant features and relevance of gender.," in *INTERSPEECH*, 2014, pp. 1248–1252.

[14] Alghowinem, S., R. Goecke, et al., "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech," in *FLAIRS*, 2012.

[15] Hussenbocus, A. and M. Lech, "Statistical differences in speech acoustics of major depressed and non-depressed adolescents," in *ICSPCS*, 2015, pp. 1–7.

[16] Scherer, S., G. Stratou, et al., "Automatic Nonverbal Behavior Indicators of Depression and PTSD : Exploring Gender Differences," in *Humaine Association Conference on ACII*, 2013.

[17] Scherer, S., G. Stratou, et al., "Automatic Behavior Descriptors for Psychological Disorders," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[18] Troisi, A. and A. Moles, "Gender differences in depression: an ethological study of nonverbal behavior during interviews," *Journal of Psychiatric Research*, vol. 33, no. 3, pp. 243–250, 1999.

[19] Williamson, J. R., T. F. Quatieri, et al., "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 4th ACM International Workshop on AVEC, ACM MM*, 2013.

[20] Williamson, J., T. Quatieri, et al., "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on AVEC,* 2014.

[21] Williamson, J. R., D. W. Bliss, et al., "Seizure prediction using EEG spatiotemporal correlation structure.," *Epilepsy & behavior : E&B*, vol. 25, no. 2, pp. 230–8, 2012.

[22] Ringeval, F., A. Sonderegger, et al., "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).*, 2013, pp. 1–8.

[23] Sethu, V., E. Ambikairajah, et al., "Phonetic and speaker variations in automatic emotion classification," in *INTERSPEECH*, 2008.

[24] Lee, C., S. Yildirim, et al., "Emotion recognition based on phoneme classes.," in *INTERSPEECH*, 2004, pp. 889–892.

[25] Bitouk, D., R. Verma, et al., "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, pp. 613–625, 2010.

[26] Le, D. and E. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *ACII*, 2015, pp. 146–152.

[27] Tipping, M., "Sparse Bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[28] Nicolaou, M. a., H. Gunes, et al., "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[29] Huang, Z., T. Dang, et al., "An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction," in *Proceedings of the 5th International Workshop on AVEC, ACM MM*, 2015.

[30] Grimm, M., K. Kroschel, et al., "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[31] Vogt, T. and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.

[32] Xia, R., J. Deng, et al., "Modeling gender information for emotion recognition using denoising autoencoder," in *ICASSP*, 2014.

[33] Paliwal, K., "Spectral subband centroid features for speech recognition," in *ICASSP*, 1998, pp. 617–620.

[34] Eyben, F., K. Scherer, et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[35] Eyben, F., F. Weninger, et al., "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[36] Cummins, N., "Automatic Assessment of Depression from Speech: Paralinguistic Analysis, Modelling and Machine Learning," *PhD Thesis*, 2016.